Predicting Purchase Trends using Black Friday Sales Dataset

Vibhor Sharma, *17BCE2152*, Anand Gopalakrishnan, *17BCE2229*, Dhruv Mittal, *17BCE2110*

***Abstract*—Every year, Black Friday happens on the fourth Thursday of November. Most products are marked down with discounts, this event is therefore a huge attraction for people and people will definitely start purchasing more. We are trying to build a model to predict Black Friday Sales through a given public dataset on Kaggle to help the shop owners stop the problem of controlling the crowd with by targeting prospective customers. To predict the sales we are going to use some machine learning algorithms like Random Forest, Decision Trees, Linear Regression, XGBoost and so on, and decide which one is the most successful.**

***Index Terms*—Black Friday Sales, Machine Learning, Neural network, Prediction model, Regression, Rule Based Learning**

I. INTRODUCTION

For huge franchises and large stores, it becomes impossible for them to know about the preferences of individual customers. Some examples of such franchises are Costco, Walmart, and Wholefoods. These stores without any proper knowledge of their customer base are struggling to satisfy the customer needs. Thus, prediction models are needed to better understand customer preferences Let us give some more context on Black Friday, it is the largest shopping day of the year in United States of America. Black Friday is the day after Thanksgiving Day which marks the beginning of the shopping season for Christmas. A prediction model developed for Black Friday can only be used during that day because customer
spending differs drastically between a normal day and a Black Friday; this is because discounts and price reductions attract more customers. Finally, better visualization techniques are required to portray the findings and help the store owners understand their customers. Our dataset is the Black Friday Sales Dataset in Kaggle. In this dataset we have the information about the Age, Occupation, City, Duration stayed, Marital status, the quantity of products bought of various types and the total amount spent. We are using these inputs to find the most necessary attributes, potentially excluding some attributes. Finally, we arrive at the conclusion from applying these models to find which model is best suited to predict Purchase trend of customers.

II. LITERATURE SURVEY

Author	**Title**	**Purpose**	**Pros**	**Cons**
Ramraj Santhanam,Nishant Uzir, Sunil Raman, Shatadeep Banarjee[2017]	Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets	Gradient boosting algorithm was developed for very high predictive capability. Still its adoption was very limited because the algorithm requires one decision tree to be created at a time in order to minimize the errors of all previous trees in the model. So it took a large amount of time to train even those models that were small in size.	In case of speed of execution, XGboost has, almost always, been seen to be superior due to the application of multithreading.	The datasets that were chosen for the experiment did not have any missing data. So, the same results cannot be expected with a dataset containing missing values and noisy information.

		Then came a new algorithm called eXtreme Gradient Boosting (XGBoost) which changed the way gradient boosting was done. In XGBoost, individual trees are created using multiple cores and data is organized in order to minimize the lookup times.		
Himani Sharma, Sunil Kumar[2017]	**A Survey on Decision Tree Algorithms of Classification in Data Mining**	As the computer technology and computer network technology are developing, the amount of data in information industry is getting higher and higher. It is necessary to analyse this large amount of data and extract useful knowledge from it. Process of extracting the useful knowledge from huge set of incomplete, noisy, fuzzy and random data is called data mining. Decision tree classification technique is one of the most popular data mining techniques. In decision tree divide and conquer technique is used as basic learning strategy. A decision tree is a structure that includes a root node, branches, and leaf nodes.	Handling each attribute with different cost.	They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal **decision tree**.
WeiWei Lin, Ziming Wu, Longxin Lin, Angzan Wen[2017]	**An Ensemble Random Forest Algorithm for Insurance Big Data Analysis**	Due to the imbalanced distribution of business data, missing of user features and many other reasons, directly using big data techniques on realistic business data tends to	The ensemble random forest algorithm is more suitable in the insurance product recommendation or potential customer analysis than traditional strong	Algorithms like SVM and LR etc. is useless in the classification of imbalance distribution feature dataset

		deviate from the business goals. It is difficult to model the insurance business data by classification algorithms like Logistic Regression and SVM etc. In this paper, we exploit a heuristic bootstrap sampling approach combined with the ensemble learning algorithm on the large-scale insurance business data mining, and proposes an ensemble random forest algorithm which used the parallel computing capability and memory-cache mechanism optimized by Spark.	classifier like SVM and Logistic Regression etc.	
Bedariani Shyti, Dhurata Valera[2018]	**The Regression Model for the Statistical Analysis of Albanian Economy**	In practice, as well as in economic theory, fulfilling strategic, managerial and operational decisions is the ultimate goal of any company. Economic performance should be assessed on the basis of the effectiveness and efficiency of resource use. The purpose of this paper is to study economic phenomena through a statistical standpoint. Our motive is to emphasize the validity of regression analysis method on economic performance.	The complexity of making decisions during statistical analysis of albanian economy increases the need for the statistical correlation method (SCM) to explain the relationship between the variables.	In this economic situation, the product incomes are the results of the conjugation of many influencing variables, but not all the defined ratios have the same importance, the action of some are compensating the others.

III.METHODOLOGY

Exploring Various algorithms used:

Regression:

- Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line). Further the line is extended or extrapolated, to predict values.

Decision Tree:

- Machine learning algorithms like decision tree and regression are used for developing a simple yet efficient prediction models.
- In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

XGBoost: It is an ensemble learning mode.

- Parallel and distributed computing makes learning faster which enables quicker model exploration. More importantly, XGBoost exploits out-of-core computation and enables data scientists to process hundred millions of examples on a desktop.
- However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

Random Forest

- Random forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.
- But, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.

In the proposed methodology, our system involves the application of machine learning techniques to predict the final target. This is done using the following steps:

1. **Data analysis:** In this step we just use common descriptive statistics techniques and apply them on our existing data like mean, median, standard deviation, frequency etc. Also, we'll find skew, kurtosis followed by correlation matrix with respect to Purchase values.

2. **Data pre-processing:** Data pre-processing is an essential step in the process of machine learning. It includes data cleaning and data partitioning This stage will involve removing all the NA (null) values and replacing them with some integer so that processing can be carried out. Later we can also display the unique value frequencies of all the columns and finally send this data into a modified train and test .csv files.
Because of our dataset being majority numerical in nature, we use the partitioning technique to remove the presence of unique non-numerical values and convert them to numerical.

3. **Parameter Selection:** Before selecting the models to use for the training process, we need to decide the columns/features that can be used as predictors and drop the others. This decision needs to be made on the basis on the data analysis done at the first stage of this process. For eg: when we made the correlation matrix of the dataset, we found that the column "purchase" was highly correlating with the column "Occupation". This implies the column occupation should be included in the predictors.

4. **Application of Machine Learning techniques:** The next step, is to use the modified train and test data and apply machine learning algorithms of various types as read about in the survey.

5. **Comparison using RMSE:** All the various models/algorithms have the same parameter for comparison i.e. the RMSE value (Root mean squared error). It can be defined as the difference between the predicted and the actual values. It is a parameter that when minimum, gives the most accurate model

6. **Exploring and trying to find newer approaches for better accuracy**. Now, our aim is to reduce the value of RMSE, to do this we can use various different models, and even explore further to find which model can provide an even better result.

IV. EXPERIMENTS AND RESULTS

First, let us start with the description of the dataset that we are going to utilize in his project. The columns are as follows:

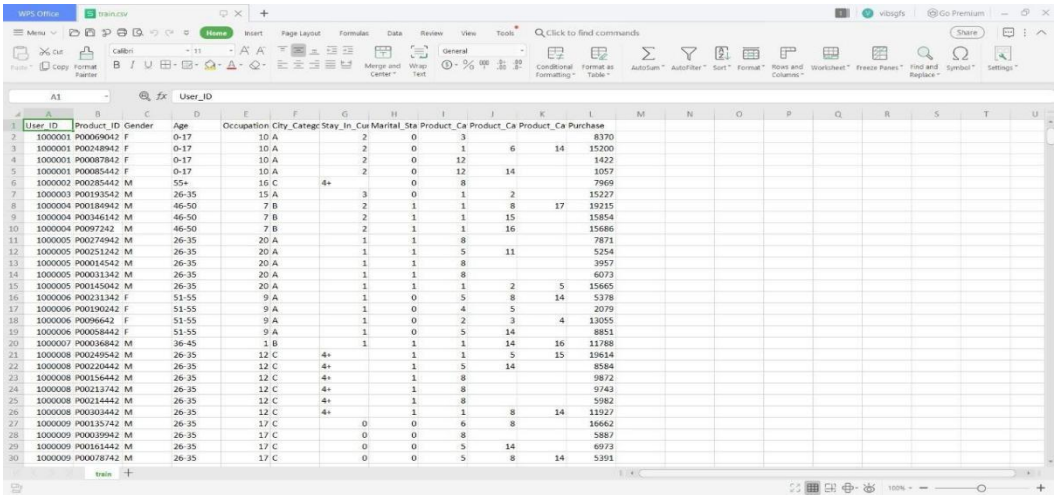
User_ID - unique
Product_ID – of the form ('P001345')
Gender – M/F (Bool)
Age - Range
Occupation – Ranging from 1-20
City_Category – A/B/C
Stay_In_Current_City_Years – Range: 1-4
Marital_Status – Yes/No
Product_Category_1 – Range: 1-17
Product_Category_2 – Range: 1-17
Product_Category_3 – Range: 1-17
Purchase

The defined target value is Purchase which contains numerical values ranging form 700015000. The dataset is

already divided into two components:

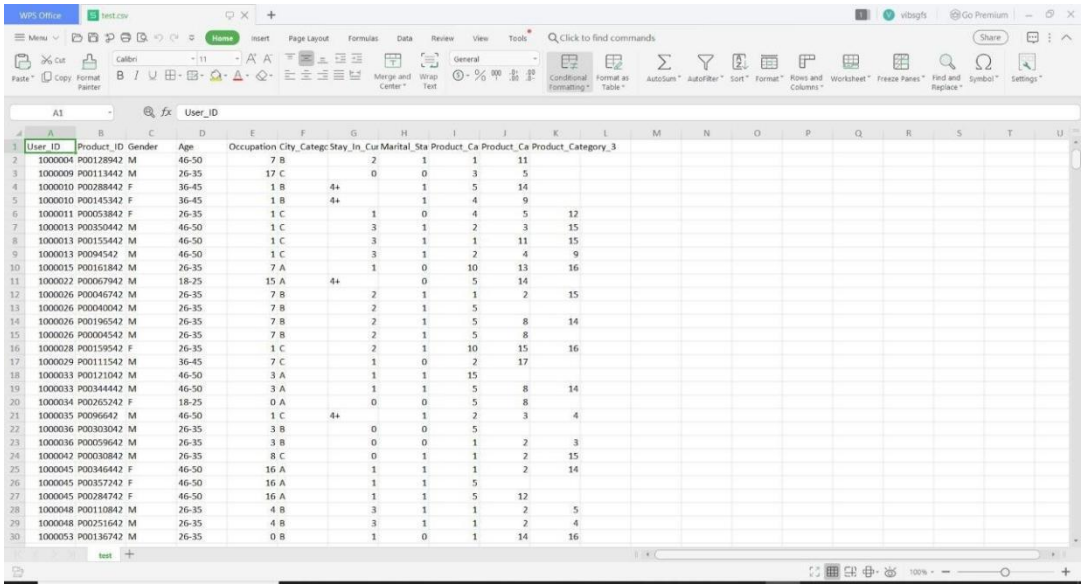
i. Train
ii. Test (No purchase values present): These purchase values need to be predicted using various models

DATASET USED:



User_ID	Product_ID	Gender	Age	Occupation	City_Categ	Stay_In_Cu	Marital_Sta	Product_Ca	Product_Ca	Product_Ca	Purchase
1000001	P00069042	F	0-17	10	A	2	0	3			8370
1000001	P00248942	F	0-17	10	A	2	0	1	6	14	15200
1000001	P00087842	F	0-17	10	A	2	0	12			1422
1000001	P00085442	F	0-17	10	A	2	0	12	14		1057
1000002	P00285442	M	55+	16	C	4+	0	8			7969
1000003	P00193542	M	26-35	15	A	3	0	1	2		15227
1000004	P00184942	M	46-50	7	B	2	1	1	8	17	19215
1000004	P00346142	M	46-50	7	B	2	1	1	15		15854
1000004	P0097242	M	46-50	7	B	2	1	1	16		15686
1000005	P00274942	M	26-35	20	A	1	1	8			7871
1000005	P00251242	M	26-35	20	A	1	1	5	11		5254
1000005	P00014542	M	26-35	20	A	1	1	8			3957
1000005	P00031342	M	26-35	20	A	1	1	8			6073
1000005	P00145042	M	26-35	20	A	1	1	1	2	5	15665
1000006	P00231342	F	51-55	9	A	1	0	5	8	14	5378
1000006	P00190242	F	51-55	9	A	1	0	4	5		2079
1000006	P0096642	F	51-55	9	A	1	0	2	3	4	13055
1000006	P00058442	F	51-55	9	A	1	0	5	14		8851
1000007	P00036842	M	36-45	1	B	1	1	1	14	16	11788
1000008	P00249542	M	26-35	12	C	4+	1	1	5	15	19614
1000008	P00220442	M	26-35	12	C	4+	1	5	14		8584
1000008	P00156442	M	26-35	12	C	4+	1	8			9872
1000008	P00213742	M	26-35	12	C	4+	1	8			9743
1000008	P00214442	M	26-35	12	C	4+	1	8			5982
1000008	P00303442	M	26-35	12	C	4+	1	1	8	14	11927
1000009	P00135742	M	26-35	17	C	0	0	6	8		16662
1000009	P00039942	M	26-35	17	C	0	0	8			5887
1000009	P00161442	M	26-35	17	C	0	0	5	14		6973
1000009	P00078742	M	26-35	17	C	0	0	5	8	14	5391

Fig 1: Training Data



User_ID	Product_ID	Gender	Age	Occupation	City_Categ	Stay_In_Cur	Marital_Sta	Product_Ca	Product_Ca	Product_Category_3
1000004	P00128942	M	46-50	7	B	2	1	1	11	
1000009	P00113442	M	26-35	17	C	0	0	3	5	
1000010	P00288442	F	36-45	1	B	4+	1	5	14	
1000010	P00145342	F	36-45	1	B	4+	1	4	9	
1000011	P00053842	F	26-35	1	C	1	0	4	5	12
1000013	P00350442	M	46-50	1	C	3	1	2	3	15
1000013	P00155442	M	46-50	1	C	3	1	1	11	15
1000013	P0094542	M	46-50	1	C	3	1	2	4	9
1000015	P00161842	M	26-35	7	A	1	0	10	13	16
1000022	P00067942	M	18-25	15	A	4+	0	5	14	
1000026	P00046742	M	26-35	7	B	2	1	1	2	15
1000026	P00040042	M	26-35	7	B	2	1	5		
1000026	P00196542	M	26-35	7	B	2	1	5	8	14
1000026	P00004542	M	26-35	7	B	2	1	5	8	
1000028	P00159542	F	26-35	1	C	2	1	10	15	16
1000029	P00111542	M	36-45	7	C	1	0	2	17	
1000033	P00121042	M	46-50	3	A	1	1	15		
1000033	P00344442	M	46-50	3	A	1	1	5	8	14
1000034	P00265242	F	18-25	0	A	0	0	5	8	
1000035	P0096642	M	46-50	1	C	4+	1	2	3	4
1000036	P00303042	M	26-35	3	B	0	0	5		
1000036	P00059642	M	26-35	3	B	0	0	1	2	3
1000042	P00030842	M	26-35	8	C	0	1	1	2	15
1000045	P00346442	F	46-50	16	A	1	1	1	2	14
1000045	P00357242	F	46-50	16	A	1	1	5		
1000045	P00284742	F	46-50	16	A	1	1	5	12	
1000048	P00110842	M	26-35	4	B	3	1	1	2	5
1000048	P00251642	M	26-35	4	B	3	1	1	2	4
1000053	P00136742	M	26-35	0	B	1	0	1	14	16

Fig 2: Test Data

In the above chapter, we defined the steps that we need to follow to get to our final objective. Let us implement the steps one by one.

1. Data Analysis

This step uses basic statistics, building correlation matrices, histograms and finding skewness and kurtosis.

The correlation matrix is essential for the process of parameter selection.

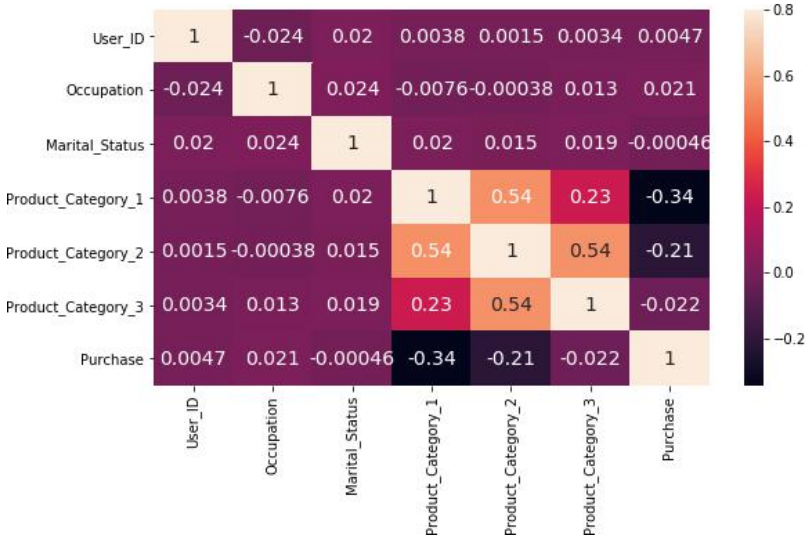
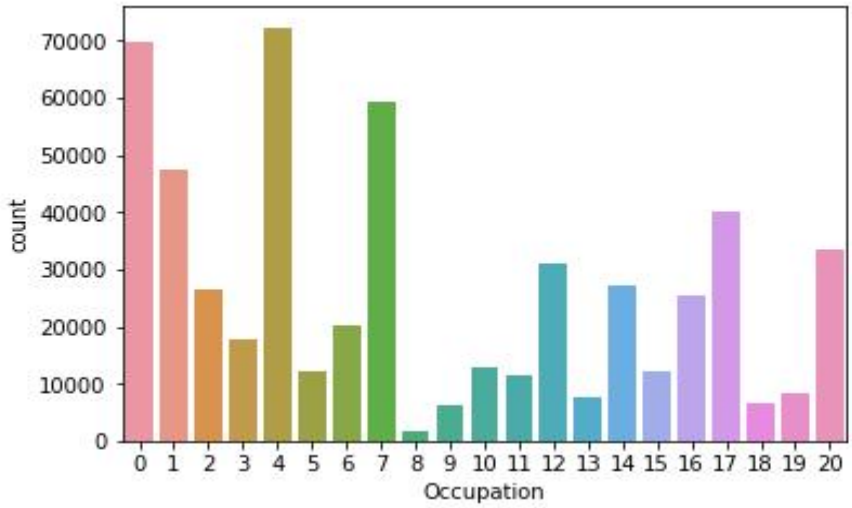
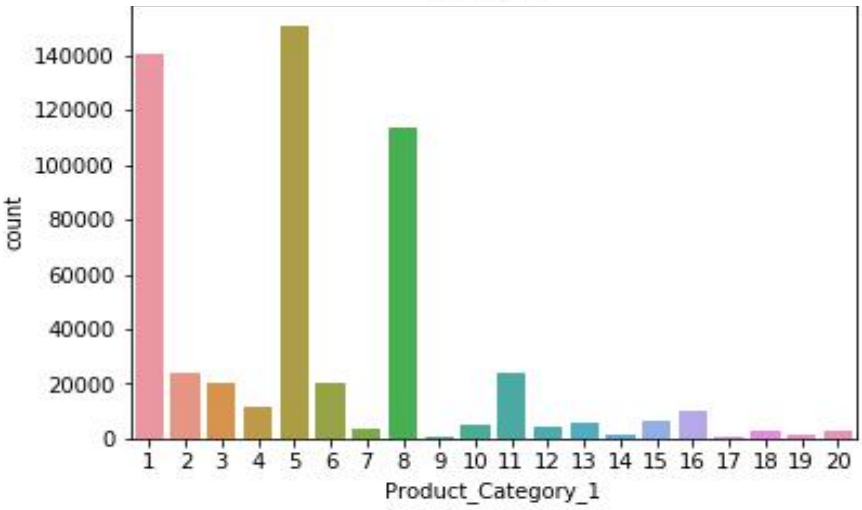
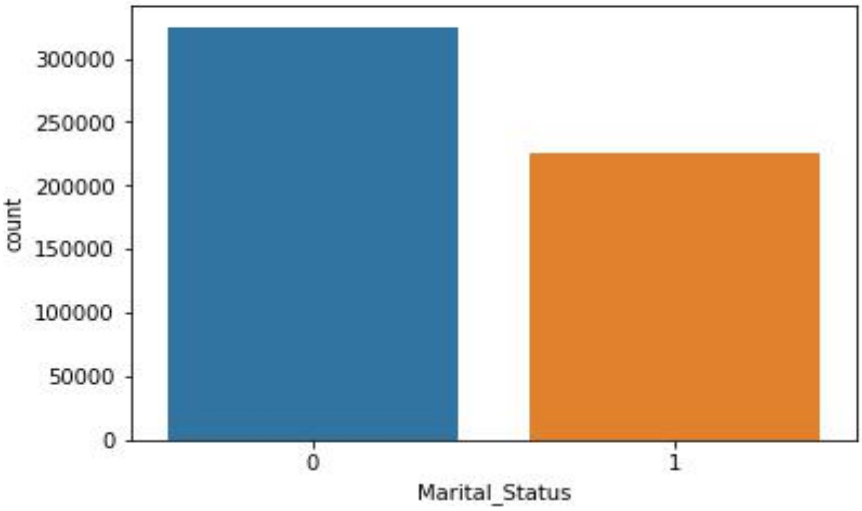
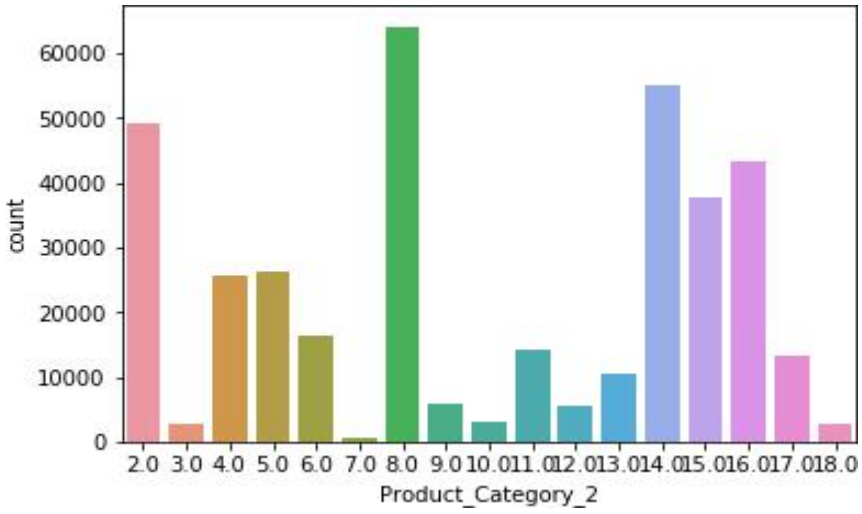


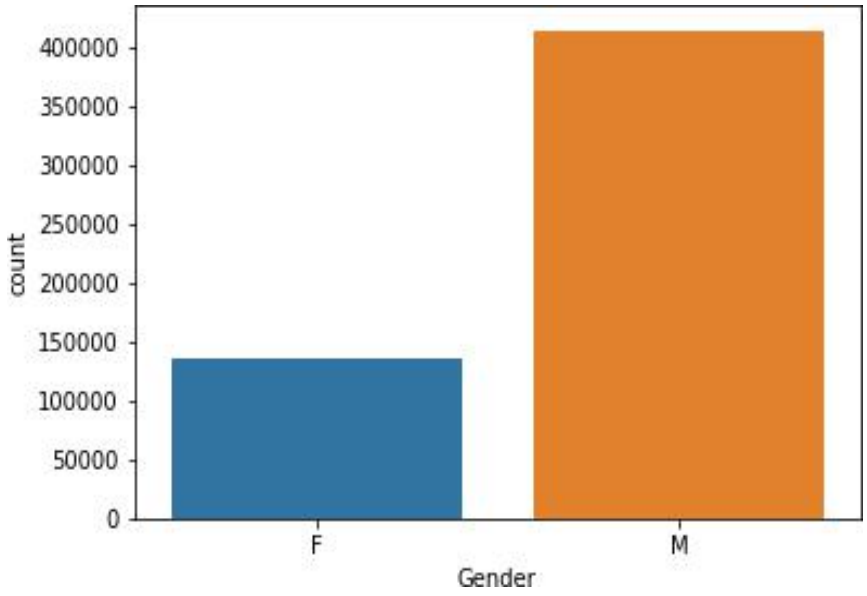
Fig 3: Correlation Matrix


70000
60000
50000
40000
30000
20000
10000
0
count
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Occupation


140000
120000
100000
80000
60000
40000
20000
0
count
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Product_Category_1


300000
250000
200000
150000
100000
50000
0
count
0
1
Marital_Status


60000
50000
40000
30000
20000
10000
0
count
2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0
Product_Category_2


400000
350000
300000
250000
200000
150000
100000
50000
0
count
F
M
Gender

2. Data Pre-processing

The Data Pre-Processing component in this project consists of the following steps:

- Data Encoding: This step removes the 'P00' part from the ProductID so that it can be used in the algorithm fitting process, otherwise a mixture of characters and numbers cannot be utilized. Hence, making it an essential step.

- Data Scaling: This process is used to reshape the data, to reduce the time that any ensemble learning algorithm might take.
- Removing Outliers: This code, allows us to drop the excess values, that exist in the columns Product_Category_1 and Product_Category_2
- Removing null values- Null and unnecessary values are removed.
- Factorizing data- Data is factorized based on the gender corresponding to their ages.
- Removing excess data- Extra data that is not not required has been dropped.

3. Parameter Selection

As stated above this process is done with the help of earlier analysis and statistics.
We selected various values from the dataset
predictors: All Columns except Purchase
target: Purchase
IDcol: User_ID, Product_ID

4. Machine Learning Algorithms

To predict the purchase amount, we implemented various machine learning algorithms and compared themon accuracy and performance metric. Since it is a regression problem, the loss function used is the Root Mean Squared error (RMSE).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

i. Linear Regression

The linear regression using python's sci-kit library was implemented on the transformed dataset. This was the simplest of the implementations in terms of complexity of the model.

Model Report
RMSE: 4632
CV Score: Mean - 4635 | Std - 35.02 | Min - 4545 | Max – 4688

ii. Ridge regression
The ridge regression using python's sci-kit library was implemented on the transformed dataset.

Model Report
RMSE: 4633
CV Score: Mean - 4636 | Std - 31.86 | Min - 4570 | Max – 4687

iii. Decision Tree Regression
Machine learning algorithms like decision tree and regression are used for developing a simple yet efficient prediction models. We used decision tree on this dataset, and it gave a good RMSE value

Model Report
RMSE: 2996
CV Score: Mean - 3242 | Std – 54.63 | Min - 3031 | Max – 3289

iv. XGBoost Model
The XGBoost is an ensemble learning model, which uses bagging and boosting. This is considered to be a highly robust model.

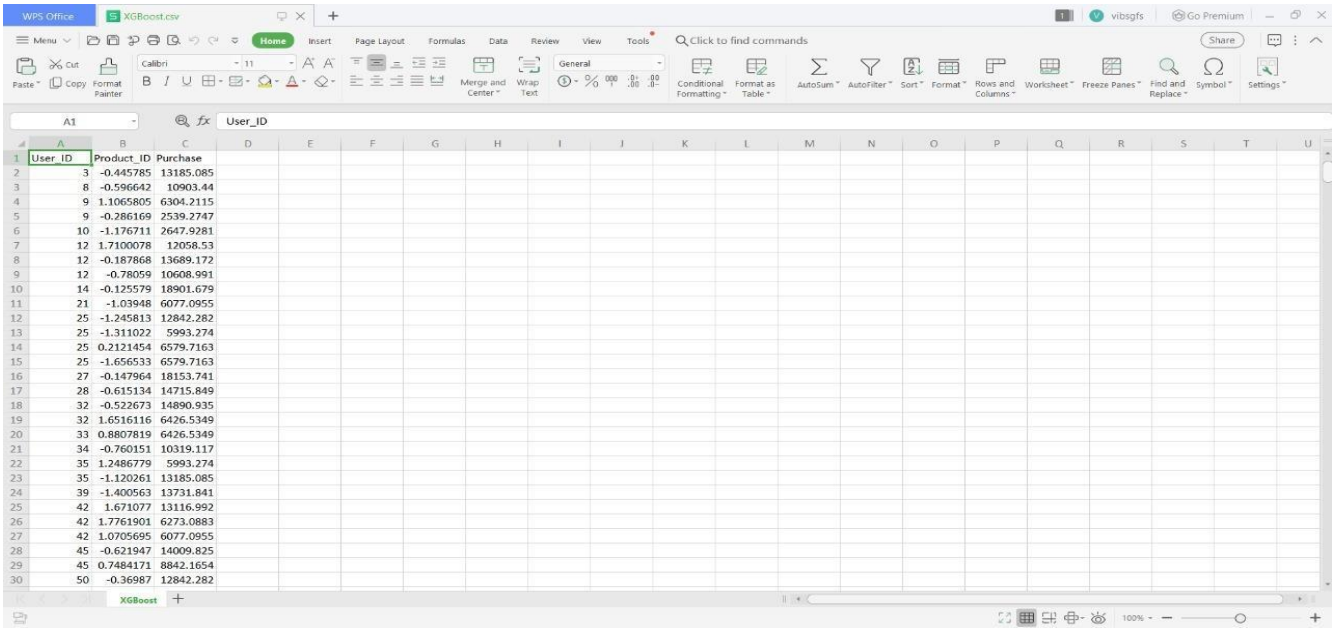
Mean Absolute Error: 240.82192676282142
RMSE: 2926

v. Random Forest Regression
Random forest Regression uses a multitude of regression Trees, and chooses the result using the best valued tree.

Model Report
RMSE: 2956
CV Score: Mean - 2973 | Std – 20.58 | Min - 2936 | Max – 3008

Hence, we can see above we implemented all the algorithms, we read about in the Literature Survey and implemented them, we got the RMSE value for each model and the features that affect it the most.

We used the model to predict the purchase values, and stored the predicted Purchase values in separate .csv /excel files.



User_ID	Product_ID	Purchase
3	-0.445785	13185.085
8	-0.596642	10903.44
9	1.1065805	6304.2115
9	-0.286169	2539.2747
10	-1.176711	2647.9281
12	1.7100078	12058.53
12	-0.187868	13689.172
12	-0.78059	10608.991
14	-0.125579	18901.679
21	-1.03948	6077.0955
25	-1.245813	12842.282
25	-1.311022	5993.274
25	0.2121454	6579.7163
25	-1.656533	6579.7163
27	-0.147964	18153.741
28	-0.615134	14715.849
32	-0.522673	14890.935
32	1.6516116	6426.5349
33	0.8807819	6426.5349
34	-0.760151	10319.117
35	1.2486779	5993.274
35	-1.120261	13185.085
39	-1.400563	13731.841
42	1.671077	13116.992
42	1.7761901	6273.0883
42	1.0705695	6077.0955
45	-0.621947	14009.825
45	0.7484171	8842.1654
50	-0.36987	12842.282

Fig 4: Excel file that contain the predicted values.

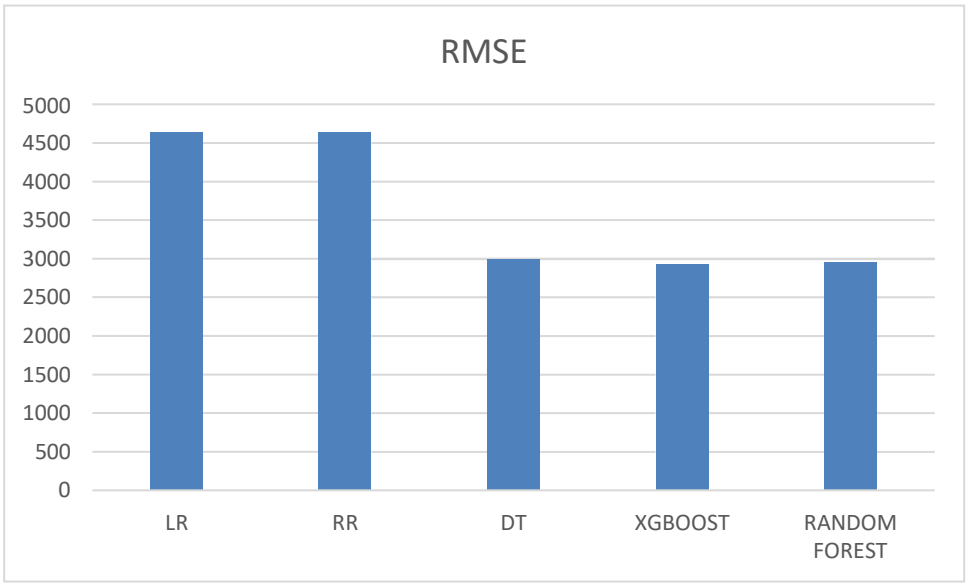


Fig 5: Tabulating all the RMSE values

Finally, we can compare all the RMSE values that we got form all the models, and we can easily see that XGBoost approach was the best one.

The predicted values form XGBoost closely matched to the already present values.

Hence proving our model was a successful one.

V. Conclusion

We can conclude by saying that we were able to complete all the objectives listed above successfully.
We applied XGBOOST which gave us a minimum RMSE value of 2926.
We were successfully able to predict the Purchase values and trend that was our primary objective, the predicted values are stored in .csv file for later consumption.

References

[1] Ramraj Santhanam,Nishant Uzir, Sunil Raman, Shatadeep Banarjee, Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets. [2017]

[2] Himani Sharma, Sunil Kumar, A Survey on Decision Tree Algorithms of Classification in Data Mining [2017]

[3] WeiWei Lin, Ziming Wu, Longxin Lin, Angzan Wen, An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. [2017]

[4] Bedariani Shyti, Dhurata Valera, The Regression Model for the Statistical Analysis of Albanian Economy. [2018]

[5] L. Bing and S. Yuliang, "Prediction ofuser 's purchase intention based on machine learning," 3rd International Conference on Soft Computing Machine Intelligence (ISCMI)., pp.99·103, Nov. 2016 .

[6] K. Singh and R. Wajgi, "Data analysis and visualization of sales data," 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), Coimbatore, pp. 1-6, Mar. 2016 .